Analysis and Synthesis of Intonation using the Tilt Model

Paul Taylor

Centre for Speech Technology Research,
University of Edinburgh,
80 South Bridge Edinburgh EH1 1HN
email
Paul.Taylor@ed.ac.uk

Abstract

This paper introduces the *tilt* intonational model and describes how this model can be used to automatically analyse and synthesize intonation. In the model, intonation is represented as a linear sequence of events, which can be pitch accents or boundary tones. Each event is characterised by continuous parameters representing amplitude, duration and tilt (a measure of the shape of the event). The paper describes a event detector, in effect an intonational recognition system, which produces a transcription of an utterance's intonation. The features and parameters of the event detector are discussed and performance figures are shown on a variety of read and spontaneous speaker independent conversational speech databases. Given the event locations, algorithms are described which produce an automatic analysis of each event in terms of the Tilt parameters. Synthesis algorithms are also presented which generate F0 contours from Tilt representations. The accuracy of these is shown by comparing synthetic F0 contours to real F0 contours. The paper concludes with an extensive discussion on linguistic representations of intonation and gives evidence that the Tilt model goes a long to way to satisfying the desired goals of such a representation in that its has the right number of degrees of freedom to be able to describe and synthesize intonation accurately.

1 Introduction

1.1 Robust Intonational Models for Speech Technology Applications

This paper presents a phonetic model of intonation designed specifically to facilitate robust computational analysis and synthesis. While intonational models of various types have been used in text-to-speech (TTS) synthesis for some time, intonation is still typically ignored completely in automatic speech recognition (ASR) systems (Granstrom, 1997). Some studies have shown uses for intonation and prosody in ASR systems (Lea, 1980), (Waibel, 1986), but it is usually the case that these components rarely make up part of state of the art large vocabulary ASR systems. The two most commonly cited reasons for the absence of intonation in ASR systems are:

1. Intonation is not a mature field and much more basic research is needed studying the phonetics and linguistics of intonation before we can apply this knowledge. Specifically, we need to discover a sophisticated and universal intonation model before applications that use such a model can be built.
2. Intonation has many functions in language, such as helping syntactic disambiguation, distinguishing new/given information, signifying word emphasis, identifying speech acts etc. None of these alone is significant enough to merit the redesign of an ASR system. In other words, it would take a lot of effort to include a specific intonational component in a recogniser and not much benefit would ensue from its inclusion.

While more basic research will certainly help the development of intonation applications, we do not think this is the main reason for the absence of intonation components in speech recognizers. In a typical contemporary ASR system (Waibel et al., 1996), (Gauvain et al., 1996), (Woodland et al., 1995), the phonetics are modelled by hidden Markov models (HMMS) and the grammar is modelled by a n-gram language model. Neither HMMs nor n-grams are a particularly "good" model of phonetics or grammar and it is known that there are many phenomena in the respective domains that neither can model. Furthermore, the use of HMMs and n-grams has not arisen through phonetic/linguistic studies which advocated their suitability: HMMs and ngrams are used because they provide simple and robust techniques for modelling their domains. Crucially they are amenable to automatic training and because of their statistical nature allow principled ways of smoothing, interpolation, merging, analysis etc. It is our belief that the main reason preventing intonation being used in ASR systems is the lack of an equivalent model for the intonational domain. In other words we disagree with statement 1 above, and argue that instead of it being fundamental research that is back holding the application of intonation, it is the lack of a suitable model which is robust, easily trainable, and amenable to statistical interpretation.

The response to the second point stems from the response to the first. For sake of argument, let us suppose that a 5% relative decrease in word error rate could be achieved if ways were found to use the above cited functions of intonation in an ASR system. If an ASR system builder had to adopt a separate approach for each of these the addition in complexity to the overall system would probably be deemed to be too great a cost for the potential increase in performance. If on the other hand a single robust intonation model could provide the basic information needed to harness all these functions, it would reduce the cost and may swing the balance in favour of using the intonational information.

While other speech technology applications such as TTS have long made use of more traditional intonational models, we believe that these applications can also benefit from the provision of a robust intonational model. In the past TTS systems typically had just a single "voice". Recently much attention has been given to the notion of having large numbers of voices in synthesis systems (**?**). A logistic requirement of this is that the speech on which these voices are modelled should be acquired quickly which implies automatic transcription techniques for all components including intonation. Hence we need some way to automatically analyse and parameterise data so that the intonational characteristics of a speaker can be captured.

1.2 Requirements of a Intonational Model

The basic aim of intonation models is to provide a system of intonational description that is linguistically meaningful in such a way that representations in this system can be automatically derived from the relevant parts of an utterance's acoustics, and that the acoustics can be automatically synthesized from the representation.

By "linguistically meaningful" we mean a representation which contains information which is significant to the linguistic interpretation of an utterance's intonation. This excludes effects which are purely redundant, or phenomena which affect the F0 contour but which are not important in this sense (e.g. segmental perturbations). Existing linguistic representations range from relatively low level phonetic descriptions such as the Fujisaki model (Fujisaki and Ohno, 1997), the Hirst model (Hirst, 1992), and the RFC model (Taylor, 1995), to higher level systems such as the IPO model (t'Hart and Collier, 1975), to phonological systems such as Pierrehumbert's (Pierrehumbert, 1980), Ladd's (Ladd, 1996) and ToBI (Silverman et al., 1992). A full discussion of the issue of linguistic representation is given in section 7, but we will now give the main desired properties of such a representation:

1. Constrained. The representation should be as compact as possible having few degrees of freedom. Specifically redundancy should be absent so that one part of the representation cannot be derived from another.

2. Wide coverage. The representation should cover as many intonational phenomena as possible and should be capable of expressing distinctions in utterances which are perceptually different.

3. Linguistically Meaningful. The form of the representation should be such that its parameters can be interpreted and generated by higher level components.

It is clear that these properties interact: an unconstrained system with many degrees of freedom will have wider coverage than a system with few. The notion of providing constrained, compact models is common throughout linguistics and it is a general rule of thumb that a compact representation system with low redundancy and an orthogonal description space is a better linguistic representation than one without these properties. Furthermore the properties of the linguistic representation interact with the two further goals, i.e.

4. Automatic synthesis. The model should have an automatic mechanism for generating F0 contours from the linguistic representation.
5. Automatic analysis. It should be possible to derive the linguistic representation automatically from an utterance's acoustics.

It is fairly easy to design a representation which is a amenable to automatic analysis and synthesis if one is not worried about the linguistic relevance of the representation.

Given the interaction between these desires, we have developed the a model of intonation that tackles these problems together in an attempt to provide a reasonable balance between them. The *Tilt model* provides a linguistic representation which is compact, has wide coverage and is linguistically meaningful. Importantly, the model has been specifically designed to facilitate automatic analysis and synthesis. The following sections now describe the representation system and the analysis and synthesis systems which allow mappings between the representation and the F0 contour.

2 Overview of the Model

The basic unit in the tilt model is the *intonational event*. Events occur as instants with nothing between them, as opposed to segmental based phenomena where units occur in a contiguous sequence. The basic types of intonational event are *pitch accents* and (following the popular terminology) *boundary tones*. Pitch accents (denoted by the letter **a**) are F0 excursions associated with syllables which are used by the speaker to give some degree of emphasis to a particular word or syllable. In the tilt model, boundary tones (**b**) are rising F0 excursions which occur at the edges of intonational phrases and as well as giving the hearer a cue as to the end of the phrase can also signal effects such as continuation and questioning. A combination event **ab** occurs when a pitch accent and boundary tone occur so close to one another that only a single pitch movement is observed. There are different kinds of pitch accents and boundary tones: the choice of pitch accent and boundary tone allows the speaker to produce different global intonational tunes which can indicate questions, statements, moods etc to the hearer.

The tilt model can be regarded as a *phonetic* model of intonation in that it describes the intonational phenomena observable in an F0 contour. This contrasts with a *phonological* model which is concerned with underlying structure of the intonation. It is only in a few practical cases that this distinction actually matters much, for example with the treatment of "level accents". These are pitch accents which no observable F0 behaviour and hence should be present in a phonological transcription but not a phonetic one.

The sequence of events in an utterance is called an *intonational stream*. A full intonational description is obtained by joining the intonational stream to the *segmental stream* (the sequence of phones) for the utterance. Bidirectional links can exists between units in one stream and units in the other stream, with the restriction that links cannot cross. Events are linked to syllabic nuclei (usually vowels), as shown in figure 1. In this way the intonation stream and the segment stream can be analysed separately and one can still find out whether a particular intonational unit is linked to a particular segment or syllable. In generative phonology such descriptions are called *autosegmental* diagrams consisting of tiers (streams) and association lines. Viewing intonation in this way is useful in that one can decouple the intonation part from the segmental part and thus compare intonation descriptions independent of the actual text they

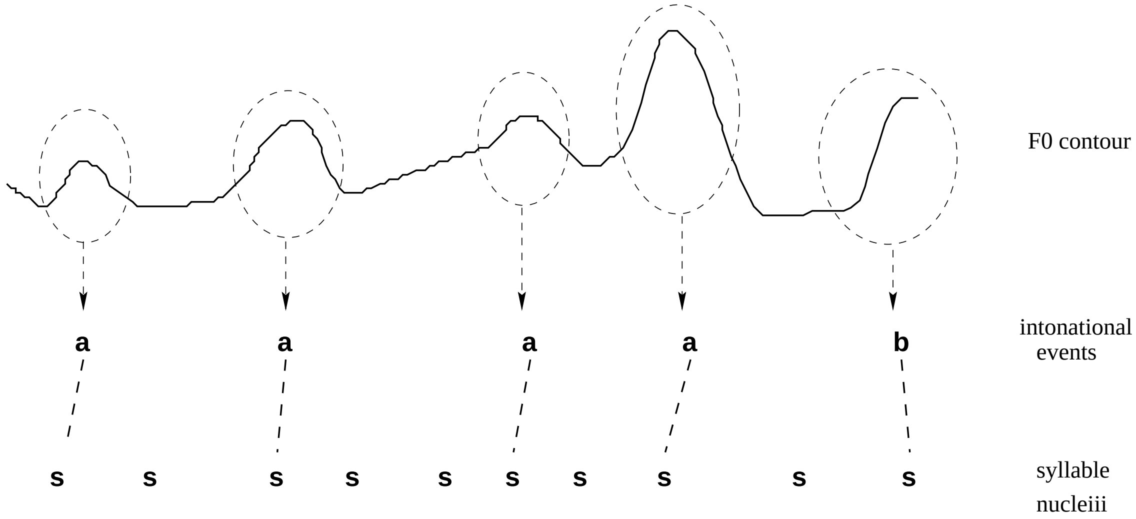

Figure 1: Schematic representation of F0, intonational event stream and segment stream in the Tilt model. The linguistically relevant parts of the F0 contour, which correspond to intonational events, are circled. The events, labelled **a** for pitch accent and **b** for boundary are linked to the syllable nuclei of the syllable stream. Note that every event is linked to a syllable, but some syllables do not have events.

are associated with. There is no loss of descriptive power in this; one can still ask whether a syllable or segment is “accented” or not.

Unlike traditional intonational phonology schemes (Silverman et al., 1992), (Pierrehumbert, 1980) which impose a categorical classification on events, we use a set of continuous parameters. These parameters, collectively known as *tilt parameters* are determined from examination of the local shape of the event’s F0 contour. A previous paper (Taylor, 1995) presented the rise/fall/connection (RFC) model. In this model, each event is fully described by a rise shape, a fall shape or a rise shape followed by a fall shape. Each event is parameterised by measuring the amplitudes and durations of the rises and falls which can be done by hand or by the curve fitting algorithm described in section 5.1. For a rise-fall shape, three points are defined which correspond to the start of the event, the peak (the highest point) and the end of the event. The rise duration is the distance from the start of the event to the peak, and the fall duration is the distance from the peak to the end; likewise, the rise duration is the difference in F0 between the F0 value at the peak and at the start, and the fall duration is the F0 distance from the end to the peak. (Hence rise amplitudes are always positive and fall durations are always negative). In this way each event is characterised by 4 parameters: rise amplitude, rise duration, fall amplitude and fall duration. If an event has only a rise component, its fall amplitude and duration are set to 0. Likewise when an accent only has a fall. The sections of contour between events are called *connections* (denoted **c**) and are also described by a amplitude and duration. (The connection is described further in section 5.3).

While the RFC model can accurately describe F0 contours, the mechanism is not ideal in that the RFC parameters for each contour are not as easy to interpret and manipulate as one might like. For instance there are two amplitude parameters for each event, when it would make sense to have only one. The *Tilt* representations helps solve these problems by transforming the four RFC parameters into three Tilt parameters, namely*duration*, *amplitude* and *tilt* itself. Duration is simply the sum of the rise and fall durations. Amplitude is the sum of the magnitudes of the rise and fall amplitudes. The tilt parameter is a dimensionless number which expresses the overall *shape* of the event, independent of its amplitude or duration. It is calculated by taking the ratio of the differences and sums of the rise and fall amplitudes and durations, as explained in section 5.2. The tilt representation is superior to the RFC representation in that it has fewer parameters without significant loss of accuracy. Importantly, it can be argued that the tilt parameters are more linguistically meaningful.

Sections 4 and 5 explain how the boundaries of events can be located from and utterance's acoustics and how automatic RFC and Tilt analysis is performed. Section 6 describes how F0 contours can be synthesised from Tilt representations. The paper concludes with a discussion on the concept of linguistic meaningfulness in intonation and its implications for the Tilt and other models.

3 Data

The three databases used in the experiments are briefly described below. Further technical details about the corpora and availability can be found in Appendix B.

3.1 DCIEM Maptask

This is a corpus of 216 dialogues collected by Canada's Defence and Civil Institute of Environmental Medicine (DCIEM)(Bard et al., 1995). Each dialogue consists of recordings of two participants playing a game called the maptask, where one participant describes a route on a map to the other participant. The maps are designed to be confusing, with the aim of illiciting interesting dialogue structures from the participants. The speech is fully spontaneous and contains many disfluencies. The database has a particularly rich variety of types of utterance, e.g. it contains many questions, instructions, statements, confirmations back-channels etc. A subset of 25 dialogues (about 2 hours of speech) was used here. Two partitions of the corpus were used. The first is a speaker independent set and comprised 20 dialogues for training and 5 for testing with none of the speakers in the training set being in the test set. All the results reported in sections 5.5 and section 6 are on the test set from this partition. One of the speakers in the corpus set appeared in several dialogues and a speaker dependent partition containing just his speech was also used.

3.2 Boston Radio News Corpus

This is a corpus of a news stories read by professional news reader, collected at Boston University (Ostendorf et al., 1995). A subset of 34 stories of about 48 minutes of one speaker was used for experiments here.

3.3 Switchboard

Switchboard is a corpus of about 2000 spontaneous speech dialogues collected live over the US telephone network (Godfrey et al., 1992). Experiments reported here are based on a 1 hour subset, chosen (by researchers at ICSI, Berkeley) to achieve maximum acoustic and phonetic variability across the corpus. Within this hour there are about 100 different speakers from all parts of the United States. 50 minutes were used for training and 10 for testing.

3.4 Hand Labelling

The databases were hand labelled to produce intonational transcriptions. The transcriptions were produced by using an interactive speech analysis tool which displayed the waveform and F0 contour, and allowed the labellers to listen to the speech. The labellers were instructed to locate pitch accents and boundaries within each utterance, in accordance with the intonational event model described above. A few extra features were added to make the labelling easier from a human point of view and to help in the error analysis of the automatic system:

- Level accents give the perception of accentuation but which have no discernible f0 movement associated with them. Although we previously said that these should not be part of a phonetic description of intonation, these were marked in the database as normal pitch accents with a **l** diacritic. The diacritic marking allows these to be ignored at a later stage if desired.
- One of the biggest problems in hand labelling intonation is that there are a large number of cases, especially in pre-nuclear position, where there is a "hint" of a pitch accent, but it is difficult to tell with certainty whether it is actually there or not. Labellers marked these as being accented and gave them a separate diacritic indicating they were "minor" accents.
- In the tilt model, only rising boundaries are classed as events. Falling boundaries are the default case and are not classified as true events. However, when labelling the corpus it was decided to give the labels **rb** to rising boundaries and **fb** to falling boundaries, again with the idea that the **fb** labels could be ignored later if desired. Normal well-formed utterances always end in one of these two labels. In spontaneous speech however many utterances are abandoned and hence it is possible for utterances to end with no boundary event.
- Silence was labelled **sil**.

3.5 Labelling Consistency

In assessing any labelling scheme it is important to give consistency figures. As well as demonstrating the inherent reliability of the task they also serve to set an upper reasonable limit of performance for an automatic system. As the automatic event detector described below is tested against human transcriptions, it is important to know how many errors in the human transcriptions we can expect.

Five labellers were used, all of whom were Edinburgh University PhD students studying various intonation topics. For comparison purposes, each of the labellers transcribed the same DCIEM dialogue. Their transcriptions were compared using a modified form of the dynamic programming scoring algorithm that is standard in the speech recognition field (see (Young et al., 1996) for an explanation). This scoring algorithm produces 2 figures, % correct which gives the total number of events correctly identified and % accuracy which is % correct minus the percentage of false insertions. The standard algorithm is modified to penalise situations where the correct label sequence is present but the timings are wrong. In intonation transcriptions, because of the small number of labels, there is a quite high probability that two label sequences will match by chance. To ensure that this isn't taken as correct, a further constraint is enforced whereby labels have to have a temporal overlap of 50% to be considered the same.

The pairwise scores for all the labellers were 81.6% correct with 60.4% accuracy. When ignoring the accents marked with the minor diacritic, the agreement is 88.6% correct with 74.8% accuracy, showing that a large number of errors were caused by minor accents. Looking at the types of events separately, the agreement for pitch accents is 81.6% correct 58.1% accuracy and the agreement for boundaries is 83.3% correct and 64.1% accuracy.

4 Automatic Detection of Events

This section describes the first stage of the automatic analysis process, namely determining the approximate event start and end positions as mentioned in section . The second stage, whereby events are assigned tilt parameters, is discussed in section 5.

4.1 Detection vs Classification

This section describes an *intonational event detector* which locates intonational events from the acoustic information alone. It is important to note the distinction between this type of system and an *intonational classifier* which use a linguistic segmentation of the utterance to perform intonational analysis. An intonational event detector is analogous to a speech recogniser in that it determines a sequence of linguistic units (words or phones in the case of a speech recogniser, pitch accents and boundaries for a intonational events) from the acoustic input alone. An intonational classifier, on the other hand, starts with a linguistic segmentation (for purposes of discussion we will assume these are syllables, but words and phones are also possible) and performs a *classification* to determine which one of N intonational categories (including unaccented) a linguistic unit has. Each approach has strengths and weaknesses which we will now briefly discuss.

Intonational classifiers have an easier task in some sense because once given a linguistic segmentation, much of the work has already been done. However, in a fully automatic system the linguistic segmentation must be done automatically also, and in certain situations (e.g. when recognising Switchboard data) the linguistic segmentations can be very error prone. This may weaken classification performance considerably. The systems also differ with respect to alignment and association. The temporal relationship between a pitch accent and its associated syllable is not simple. Experiments with the tilt model have shown that only about 50% of accents have their peak within the boundaries of the associated syllable: the remainder have either late or early peaks which are actually closer to adjacent syllables. Event detectors show the precise location of events in time but do not (in the first instance) show which syllable or word is accented. Conversely, classifiers show that a certain syllable is accented or unaccented, but do not say where in relation to the syllable the accent is to be found.

The choice of which approach to take is based on the what the intonational analysis system is to be used for. Specifically it depends on whether it is reasonable to assume that a linguistic segmentation is actually available, and which is more important: knowing the precise location of the accent or knowing which syllables are accented. Here we report an automatic intonational event detector; work by others has already been performed on intonational classifiers (e.g (Ross and Ostendorf, 1995)). The tilt model itself can work with either approach; all it needs is the approximate location of the events, which both approaches can provide.

A final point concerns accuracy measurements of the systems. It is important to note that it is not possible to meaningfully compare accuracy figures for the two types of system. The accuracy figures for classification systems can always be expected to be considerably higher for two reasons. Firstly, for reasons of system development, the linguistic segmentation is normally assumed to be perfect and so some degradation is to be expected when used with a fully automatic system. Secondly, and more importantly, classification results normally report how well the system identified unaccented syllables. As these may typically account for 70%-80% of the syllables in the test set, it is important to see the baseline accuracy for such systems as being this figure. In event detection, there is no such baseline and because of insertion errors, figures may even be worse than 0%.

4.2 Event Detector Overview

The automatic event detector uses continuous density hidden Markov models to perform a segmentation of the input utterance. A number of units are defined and a HMM is trained on examples of that kind from a pre-labelled training corpus using the Baum-Welch algorithm (Baum, 1972). Each utterance in the corpus is acoustically processed so that it can be represented by sequence of evenly space frames. Each frame is a multi-component vector representing the acoustic information for the time interval centred around the frame.

Recognition is performed by forming a network comprising the HMMs for each unit in conjunction with a n-gram language model which gives the prior probability of a sequence of n units occurring. To perform recognition on an utterance, the network is searched using the standard Viterbi algorithm to find the most likely path through the network given the input sequence of acoustic vectors.

Using the HTK toolkit (Young et al., 1996), a series of experiments were performed, each following the same experimental procedure. First a HMM set is defined, each HMM representing one intonational unit (such as **a** or **c**). Each HMM has three states, each of which initially has a single Gaussian which gives the probability density function for the acoustic data. The HMM parameters are initialised using the Viterbi algorithm to provide starting estimates for the model parameters. Training proper is performed using the Baum-Welch algorithm. A number of training iterations are performed until convergence is reached. The single Gaussian is then split into two Gaussian components to form a Gaussian mixture for that state and the Baum-Welch algorithm is run again. This process is repeated for models of 2, 3, 4, 5, 6, 7, 8, 10, 12, 14, 16, 20, 24 and 28 components. All the experiments used the non-embedded form of the Baum-Welch algorithm. In this style of training each HMM is trained on frames of speech lying within the boundaries of its own units only, as opposed to embedded style training where only the sequence of units is given and it is up to the training algorithm to assign frames appropriately. Experiments showed that this style of training consistently produced better results that embedded training.

Testing is performed by running the trained system over test data and comparing the transcriptions to those produced by a hand labeller using the procedure outlined in section 3.5. This technique simply decides if an event in the automatic transcription corresponds to an event in the reference transcription. In most of the results reported below, all events are treated as the same category, so a pitch accent in the test transcription can be successfully matched to either a pitch accent or boundary tone in the reference transcription. Section 4.4 gives individual scores for pitch accents and boundary tones. The comparison procedure gives %correct and %accuracy for the standard case and for the case where minor accents are ignored. In all cases, the %accuracy for the standard case is taken to be the most important measure, and is the one used for determining the highest scoring system.

Sections 4.3 and 4.4 report experiments on different feature and label configuration for the DCIEM test data. Section 4.5 compares these with results for the Boston Radio News Corpus and Switchboard. It should be noted that none of this data was used in the development of the RFC and Tilt models. The HMM event detector was developed on a training and development test set which are now included in the DCIEM training set, and the test sets for the corpora can properly be considered unseen evaluation data.

4.3 Features

The super-resolution pitch detection algorithm (Medan et al., 1991) was used to extract F0 contours from waveforms for the DCIEM database. This algorithm has been shown to be one of the most accurate F0 detection algorithms currently available (Bagshaw et al., 1993), but contours extracted from any state-of-the-art algorithm should be adequate for use in the model. The integrated pitch tracking algorithm (Secrest and Doddington, 1993) was used for Switchboard because it gave better results on telephone speech (which is often missing the fundamental). RMS energy was calculated in the standard way. The F0 and RMS values were combined to give a feature representation at 10ms frame intervals.

Table 1 gives the results for 4 experiments on different acoustic feature sets. Experiment F1 used plain F0 and rms energy. Experiment F2 used a simple form of speaker and channel normalisation whereby the mean and standard deviation of each speaker's F0 and energy was calculated and used to normalise all the data for that speaker.

It has been shown before (e.g. (Taylor, 1995)) that the *change* in F0 is a particularly salient cue to the presence of a pitch accent, and so the normalised F0 and energy measures were supplemented by

Features	% c	% a	% major c	% major a
F2 F0 and energy	57.7	26.6	69.6	46.3
F1 Normalised F0 and energy	61.7	33.6	73.0	51.7
F3 Normalised F0 and energy + deltas	65.6	43.8	76.7	56.1
F4 Normalised F0 and energy + deltas + acc	72.7	47.7	81.9	60.7

Table 1: Performance for different feature sets

their delta coefficients. Deltas were calculated in the standard way by taking an estimate of the first derivative of a value over a period of 4 frames. It is also possible to calculate a second order delta which gives the rate of change of the normal delta coefficients. Experiment F3 gives the results for the normalised and delta coefficients and experiment F4 gives the results for the normalised, delta and delta-delta coefficients. As feature set F4 gave the best results, its feature combination was used for all the subsequent experiments.

4.4 Labels

As far as the tilt model is concerned, there are only five intonation labels, namely **a**, **b**, **ab**, **c** and **sil**. However, for the reasons described in section 3.4, a richer label set was used for hand labelling which differentiated rising and falling boundaries, and had diacritics for level and minor accents. A series of experiments were performed to see which label set was the optimal for the HMM event detector. These experiments investigated whether level accents, minor accents and falling boundaries should be included in the label set.

Name	Labels	Description
L1	**sil, c and e**	Major pitch accents and rising boundaries are **e**. Falling boundaries, minor and level pitch accents are **c**
L2	**sil, c, a, ab, a**	Normal, minor and level pitch accents are **a**; all rising boundaries are **b**, falling boundaries are **c**
L3	**sil, c, a, ab, a**	Normal pitch accents are **a**; rising boundaries are **b**;minor and level pitch accents and falling boundaries are**c**
L4	**sil, c, a, fb, rb, afb, arb, m, mfb, mrb, l, lrb, lfb**	full label set

Table 2: Performance for different label sets

Four label sets, shown in table 2, were defined to investigate the various issues just outlined. Label set L1 is the simplest possible set, where pitch accent and boundary labels are mapped to a single label **e**, representing all events. In sets L2 and L3, rising boundaries are labelled **b**, falling boundaries are ignored (i.e. they are labelled **c**), and pitch accents are labelled **a**. In set L2, level, minor and normal pitch accents are grouped into a single accent category **a**, while in set L3, only normal accents are labelled **a**, level and minor accents are ignored (labelled **c**). In set L4, all variations are given their own label, so that level accents are labelled **l**, minor accents **m** and the combined accent and boundary labels for each (**ab** for normal accents) are also marked separately.

A separate recognition experiment was performed for each set of labels. In testing, as before, the identities of the event labels were treated as equivalent, allowing direct comparison across label sets. It is clear from the results given in table 3 that the sets with finer event distinctions out perform the sets where different types of events are grouped together. The best performing set is L4, where each possible type of event has its own HMM. Hence this label set was adopted as the standard set in the event detector and used for all the other experiments, including those previously reported on feature usage.

Labels	% c	% a	% major c	% major a
L1	60.2	43.8	73.4	56.9
L2	70.5	46.9	80.4	56.9
L3	67.6	44.1	77.9	54.4
L4	72.7	47.7	81.9	60.7

Table 3: Performance for different label sets

In all the results reported here, events of one label matching events of another are considered correct. To show individual labelling accuracy however, a set of comparisons were performed where labels had to match their own type to be considered correct. Table 4 shows that labels often don't match themselves very well, for instance when **a** accents are compared to **a** accents in the reference transcription the accuracy is only 25.9%. As expected, minor and level accents are extremely difficult to detect and have very low accuracy (-24.2% and -52.4%). However, when **a, l** and **m** accents are allowed to match with each other, the performance is substantially higher. The accuracy for boundary event detection is relatively low at 19.2%. The further results for boundary detection show that the source of errors is almost entirely due to falling boundaries **fb** being missed (-25.9%), the score for rising boundary detection is substantially higher at 34.8% accuracy. These figures tell us two things. Firstly (and not unexpectedly), events which are not distinct acoustically are detected with much lower accuracy than those which have prominent acoustic features. Secondly, although discrimination between the three accent types is poor, collectively they actually produce better accent recognition than when a single model is trained for **a, l** and **m** (Label set L2 gives 73.1% correct 40.5% accuracy for accents).

Reference Label	Test Label	% correct	% accuracy
a	a	71.9	25.9
m	m	3.4	-24.2
l	l	9.8	-52.4
a l m	a l m	70.9	44.2
fb rb	fb rb	58.0	19.2
rb	rb	55.0	34.8
fb	fb	49.1	-25.9

Table 4: Individual performance for different labels

4.5 Datasets

The above results give event detection performance on the speaker independent DCIEM test set (SI-DCIEM). Further experiments were performed on a single speaker subset of this (SD-DCIEM), the Boston University Radio News Corpus (RN) and Switchboard (SWB), all using the F4 feature set. Results are shown in table 5. The results for the SI-DCIEM corpus are the same as those in tables 1 and 3. The SD-DCIEM corpus contains about 30 minutes from a single speaker, and this was used to examine the differences between a speaker independent and speaker dependent event detector. There is a clear improvement in performance from the speaker independent to the speaker dependent test.

The three different results of the Radio News corpus correspond to different labelling situations. RN1 corresponds to the L4 label set as used in the other experiments, while RN2 and RN3 correspond

Dataset	% c	% a	% major c	% major a
SI-DCIEM	72.7	47.7	81.9	60.7
SD-DCIEM	82.1	63.1	88.1	70.2
Radio News 3	69.4	49.7	79.4	59.3
Radio News 1	68.9	49.2	n/a	n/a
Radio News 2	67.1	45.7	n/a	n/a
Switchboard	60.7	35.1	71.5	47.4

Table 5: Performance for different data sets

to transcriptions which have automatically been converted from ToBI transcriptions. The RN corpus had been previously labelled at Boston University using the ToBI scheme and we investigated whether a database already labelled with the ToBI scheme could be converted into the tilt model scheme. The mapping, which is quite complex, is described in Appendix A. There are two variants. In RN2 all ToBI boundary tones are labelled as **b**, and in RN3 **H%** tones are labelled as **rb**, and **L%** tones are labelled as **fb**. The RN corpus is from a single speaker only and so it is somewhat surprising that the results, although better than SI-DCIEM, aren't as good as SD-DCIEM. This result led us to further analysis of the speaker independent DCIEM where we examined the errors attributed to each of the 10 speakers in the test set separately. We found there was quite a large variation in performance, with the best speaker having 76.3% correct and 58.9% accuracy, and the worst 64.3% correct and 33.9% accuracy. The difference in performance between SD-DCIEM and RN lies within this range and hence the differences may be attributed to some speakers being naturally more suited to the approach than others.

It is interesting to note that the best results for the ToBI mapped labels (RN2 and RN3) are nearly the same as the results for the tilt labels (RN1). This is an important result because it means that the event detection technique described here can be used on databases already labelled with the ToBI scheme. However, although the performance of the RN1 and RN3 is similar, the actual transcriptions they produce are significantly different. The results in the table were obtained by testing the RN1 trained event detector against the test set labelled using RN1. Likewise the RN3 event detector was tested against a RN3 labelled test set. When the RN1 event detector is tested on the RN3 labelled test set, the performance drops to 49.1% correct with 40.2% accuracy. The biggest discernible difference in the two sets is that the Tilt set (RN1) has many more accent labels than the ToBI (RN3) set. Whether this is due to a fundamental difference in the labelling schemes, or just a discrepancy between labellers is difficult to say. In summary then, it is possible to map ToBI labelled data into event labels and train an accurate event detector, but one should not assume that the resulting labelling from this is the same as from a Tilt event detector.

The results for Switchboard are worse than for the other datasets. Accurate word transcription of Switchboard has proved a notoriously difficult task for speech recognition systems, which often perform much worse on this task than others. Many reasons are given for Switchboard's difficulty including disfluencies, "poor" pronunciations (i.e. substantially different from citation forms) and low acoustic quality. As regards the performance of the event detector we can probably rule out the spontaneous nature of the speech as being the source of the poorer performance. Although the task is different, the DCIEM corpus is fully spontaneous also and contains highly disfluent speech. The DCIEM corpus is recorded from speakers of a fairly homogeneous regional accent group, while switchboard is from speakers from across the entire United States and this may account for some of the worse performance. However, in our opinion it is the differences in acoustic quality which are probably the most important factor. While DCIEM was recorded with high quality close talking microphones in a quiet room, switchboard was recorded live over the US telephone network. From listening to the speech, the acoustic quality of Switchboard is very bad in places, with background noise, and telephone network artifacts. The poor acoustic quality affects feature extraction with many more errors in F0 been present than with other

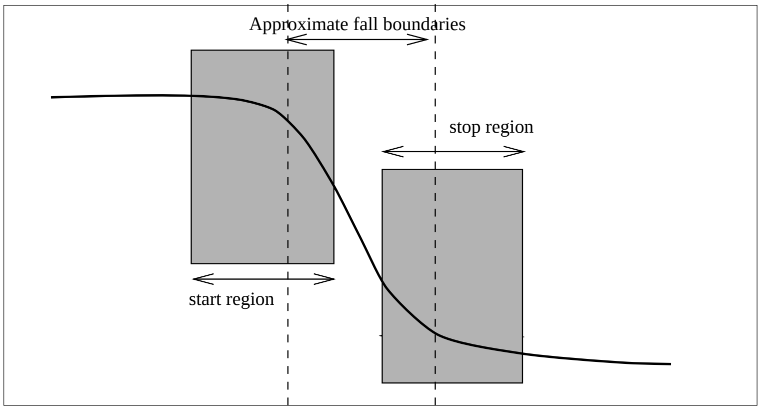

Figure 2: Search regions for fall accents

databases. However, given that Switchboard is such a difficult database, the figures of 60.7% correct and 35.1% accurate may not be too bad.

5 Deriving Tilt Parameters from Event Boundaries and F0 contours

The event detector produces a segmentation of the utterance from which it is possible to derive the start and end positions of the events. This information is used to delimit a region of contour that is first turned into RFC parameters and then Tilt parameters. Each of these processes is now described in turn.

5.1 Automatic RFC Analysis

Automatic RFC analysis involves determining the precise locations of the start, peak and end positions and using the values to calculate the rise amplitude, rise duration, fall amplitude and fall duration of the event. This process is explained in more detail in Taylor (1995); here we present a summary.

RFC analysis operates only on sections of F0 contour which have been delimited by the event detection procedure. Each of these sections is smoothed using a median smoothing algorithm, and unvoiced regions are interpolated through. This ensures that the RFC analysis sees only smooth fully voiced contours. Smoothing serves a number of purposes. Firstly median smoothing is useful for removing isolated spurious F0 values produced by the errors in the extraction process. Secondly, it helps remove the natural minor perturbations in F0 periods which result from natural variations in the speakers production. While these perturbations affect speech quality, they are not important in intonation analysis and synthesis and can be removed without distorting the intonational content of the contour. A median filter with a window of about 7-11 points is sufficient to smooth the contour.

After smoothing, a peak-picking algorithm is used to determine whether the event is a rise, fall or combined rise-fall. If a peak is found, then the event is classified as a combined rise-fall. The peak position (if present) and the start and end position as given by the event detector, are used to define *search regions*. In the case of a single rise or fall event (as shown in figure 2) the search regions are defined to be 20% before and after the approximate event detector boundaries. Typically this will correspond to ten 10ms frames for the start and ten frames at the end. Each start frame position in combination with each end frame position is taken as a potential start and end point, and a F0 curve is synthesized for each of these start and end combinations (in our example this is $10^2 = 100$ curves). Each of these curves is compared with the values of the actual F0 contour at that point and the curve with the lowest Euclidean distance is taken to be the best fit. Compound rise-fall events are treated similarly, but in this case two searches are performed. The first search (to find the rise) defines its start search region as before, but the end position is fixed as the peak. The second search (to find the fall) has a fixed start position at the peak and has a variable end search region as above. This procedure is continued until all the precise start, peak and end times have been located for every event in utterance.

5.2 Automatic Tilt Analysis

The tilt representation is easily derived from the RFC representation by application of the equations described below. The *tilt* parameter itself is an abstract description of the F0 shape of an event. Tilt is calculated from comparing the relative sizes of the amplitudes and durations of the rises and falls for an event. Amplitude tilt is given by

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{1}$$

and duration tilt is given by

$$tilt_{dur} = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}} \tag{2}$$

Empirical evidence has shown that these parameters are highly correlated (see section 5.5) to the extent that a single parameter can be used for both amplitude and durational tilt. This single value is calculated from the averages of both:

$$tilt = \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| + |A_{fall}|)} + \frac{D_{rise} - D_{fall}}{2(D_{rise} + D_{fall})} \tag{3}$$

The other two tilt parameters, amplitude and duration, are calculated in terms of the sum of the magnitudes of the rises and falls.

$$A_{event} = |A_{rise}| + |A_{fall}| \tag{4}$$

$$D_{event} = D_{rise} + D_{fall} \tag{5}$$

5.3 Segmental and Event Based Views

So far events have been described using what amounts to local information about their amplitude, duration and shape. There are a number of ways to describe further information, specifically how they are located in time with respect to the rest of the utterance.

There are two possible formally equivalent ways to do this. The first makes use of the filler unit *connection*, implying a segmental based view in which all the information is presented as a contiguous sequence of units, one ending where the next starts. In this view, connections take durations such that the total duration of an intonational description can be calculated by summing the durations of the connections and events. Connections also have amplitudes so that the starting f0 value of an event following a connection is given by the end F0 value of the previous event plus the amplitude of the connection.

An alternative view, which does away with the need for connections, is closer to the philosophy of the event based formulation described in the overview. Connections can be eliminated by explicitly attaching time and distance from baseline parameters to the events. For the F0 value we designate a parameter *start-F0* which specifies the height in Hz to the start of the event. The simplest way to specify the position parameter is with respect to the start of the utterance, for example saying that the start of the event is 2.5 seconds from the start of the utterance. An alternative is to measure position with respect to the syllable with which the event is associated. In practice we have use a measure which gives the distance from start of the nucleus of the syllable (usually the vowel) to the peak (the join between the

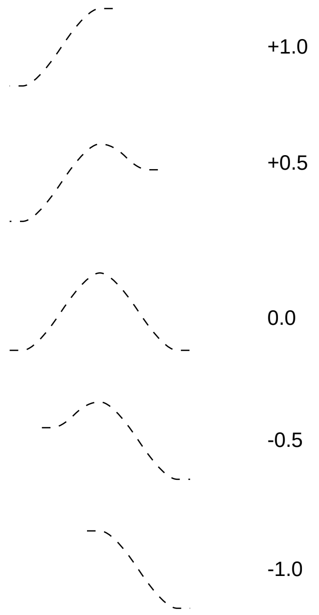

Figure 3: Examples of 5 events with varying values of tilt

rise and fall) of the event. If the event is rise only the end of the rise is used, if the event is fall only the start of the fall is used. The start of the nucleus is used because it is easy to locate in a segmented utterance and because the start of the syllable itself (that is, the boundary between the current syllable and the previous one), is often difficult to determine.

Measurement with respect to the start of the utterance (absolute position) and with respect to the associated syllable (syllabic position) each have their own advantages. Absolute position is useful when the intonation stream is produced in isolation and where the syllable stream is not present. Syllabic position is useful in that it behaves similarly to the three proper tilt parameters, and can be considered to be a local parameter which carries intonational significance.

The information in all the formulations is exactly equivalent and one can be mapped to an other without loss of information. The decision to use one rather than an other is often made on the basis of the practical application in which the model is being used.

5.4 Interpretation of Tilt Parameters

Here we briefly describe the significance of each of the Tilt parameters. Amplitude corresponds to the phonetic prominence of an event - the bigger the amplitude of an event in a given location the more prominence it will receive. Amplitude is measured using the linear Hertz scale. This has the advantage in being easy to interpret, but some additional processing is often required to obtain a more linguistically representative prominence value. For instance, pitch range often narrows towards the end of an utterance and hence an accent excursion of size xHz at the start of an utterance will be perceived as less prominent than an accent of xHz at the end (Liberman and Pierrehumbert, 1984).

Duration is measured in seconds. Compared with the other tilt parameters, duration doesn't contain much genuine high level intonational information. While it would in principle be possible to collapse amplitude and duration into a single quantity of intonational *size*, the correlation studies have shown that duration and amplitude are not highly correlated and their amalgamation would result in a substantial loss of synthesis accuracy. The variance in the duration parameter mainly arises because of the interaction between the intonation and segmental streams. Differences in duration are often a function of the size of

	rise amplitude	rise duration	fall amplitude	fall duration
rise amplitude	1.0			
rise duration	0.33	1.0		
fall amplitude	-0.48	-0.04	1.0	
fall duration	-0.18	-0.46	0.025	1.0

Table 6: Correlation matrix for RFC parameters. Values near 1.0 indicate positive correlation, values near -1.0 indicate negative correlation and values near 0.0 indicate little correlation. Note that the matrix is symmetric and hence only the bottom left corner is shown for clarity

	amplitude	duration	tilt
amplitude	1.0		
duration	0.17	1.0	
tilt	0.06	-0.09	1.0

Table 7: Correlation matrix for Tilt parameters

the voiced interval that an event can be realised within: some syllables are longer than others and crucially some have substantially more voicing. It is common to see events associated with short syllables (e.g. "pot") to have short durations, while events associated with longer syllables (e.g. "strength") to have longer duration.

Tilt is a measure of event shape and represents the relative sizes of the rise and fall components of an event. A value of +1.0 indicates a rise, a value of -1.0 indicates a fall, and a value of 0.0 indicates an event with equal sized rise and fall components. Figure 3 shows 5 different event shapes with their tilt values. Tilt is dimensionless and is not dependent on amplitude or duration.

In addition to the three core tilt parameters it is also worth mentioning the significance of the *syllabic position* parameter. As Ladd (Ladd, 1996) points out, when discussing temporal relations, it is important to distinguish *association* from *alignment*. Association describes the structural relationship between the intonation stream and the segment stream by saying which units in one are linked to units in the other. Alignment on the other hand describes the temporal relationship between units, and can be important in distinguishing pitch accent type. The syllabic position parameter is used to represent alignment in the tilt model. In rise-fall events, syllabic position is the distance between the peak of the event (i.e. the boundary between the rise and fall) and the start of the nucleus of the syllable that the event is associated with (the accented syllable). In simple rise or fall events it is the distance between the start of the event and the start of the vowel of the accented syllable. An event which has the same amplitude, duration and tilt parameters can signal different effects depending on its position. For instance, what are known as H* and H+L* accents in the Pierrehumbert notation can often be realised by a simple falling contour: the difference is that the H* occurs much later with respect to the vowel than the H+L*. Further discussion of the position parameter is given in section 7.4.

5.5 Modelling Accuracy of the Tilt and RFC models

The motivation behind mapping RFC parameters into Tilt parameters is to produce a new representation which has less redundancy and is more linguistically meaningful. An assessment of the linguistic relevance of the tilt model is left to section 7, but here we can give some evidence about the redundancy in the RFC and Tilt representations.

A useful way to examine redundancy in a set of data is to calculate its correlation matrix, which shows the correlation of every parameter against every other parameter. Table 5.5 shows the correlation matrix for the RFC parameters as measured on the DCIEM test set. We can see clearly from the table that a number of parameters are indeed highly correlated, for instance rise amplitude against fall amplitude and rise duration against fall duration.

As explained above, the tilt parameter is calculated by averaging amplitude tilt and duration tilt into a single parameter and it is this which allows the four RFC parameters to be reduced to three tilt parameters. Is this justified? By combining the two tilt parameters in equation 3 we are effectively saying that they are equal: taking the average of the two is simply more robust than using either one alone. We can rearrange the equivalence equation 6 to give equation 7

$$\frac{|A_{rise}|-|A_{fall}|}{(|A_{rise}|+|A_{fall}|)} = \frac{D_{rise}-D_{fall}}{(D_{rise}+D_{fall})} \tag{6}$$

$$\frac{|A_{rise}|}{D_{rise}} = \frac{|A_{fall}|}{D_{fall}} \tag{7}$$

which states that the magnitude of the gradient of the rise part is equal to the magnitude to the gradient of the fall part. When we actually measure the correlation of the rise and fall gradients for our data we get a high correlation of 0.64, which is higher that any shown in table 5.5. The correlation between the amplitude and duration tilt parameters themselves is 0.73. Hence collapsing the amplitude and duration tilt parameters based on correlation is justifiable. Table 5.5 shows the correlation matrix of the tilt parameters for the same data. There is a slight correlation between amplitude and duration (0.17), but virtually no correlation between tilt and amplitude and tilt and duration.

In theory, it is easy to map a set of n-dimensional parameters to a set of n-1 dimensional parameters using standard techniques such as principal component analysis. What is more difficult is to achieve parameter reduction without significant loss of information. While the correlation figures just quoted prove that the Tilt system provides a compact set of intonational control parameters which are independent of one another, it is important to prove that little information has been lost in the process. The next section explains how RFC and Tilt representations can be converted back into F0 contours and examines the accuracy of this process. Crucially, we show that the reduction of the four RFC parameters to the three tilt parameters is achieved without significant information loss.

6 Tilt Synthesis

The synthesis process of converting Tilt representations into F0 contours involves two steps: converting Tilt representations into RFC representations and then converting these into F0 contours. Given the tilt parameters for an event, the RFC parameters can be calculated by equations formed by rearranging equations 3 4 and 5:

$$A_{rise} = \frac{A_{event}(1+tilt)}{2} \tag{8}$$

$$A_{fall} = \frac{A_{event}(1-tilt)}{2} \tag{9}$$

$$D_{rise} = \frac{D_{event}(1+tilt)}{2} \tag{10}$$

$$D_{fall} = \frac{D_{event}(1-tilt)}{2} \tag{11}$$

The conversion process first involves converting event style descriptions into segmental style descriptions (that is *start F0* and *position* parameters are converted into connection information). Next, equations 8-11 are used to produce the RFC parameters for the events. Each event is decomposed into its separate rise and fall components, each of which is synthesized using the following equation:

$$\begin{array}{ll} f_0(t) = A_{abs} + A - 2.A.(t/D)^2 & 0 < t < D/2 \\ f_0(t) = A_{abs} + 2.A.(1 - t/D)^2 & D/2 < t < D \end{array} \qquad (12)$$

where A is rise or fall amplitude, D is rise or fall duration and A_{abs} is the absolute F0 value at the start of the rise or fall, which is given by the end value of the previous event or connection. Connections are synthesised using straight lines:

$$f_0(t) = A_{abs} + A.(t/D) \quad 0 < t < D \qquad (13)$$

where A is connection amplitude, D is connection duration and A_{abs} is as before.

6.1 Synthesis Accuracy

We now address the question of synthesis accuracy. This is measured by taking a tilt representation for an utterance in the data, synthesizing an F0 contour from this and measuring the difference between the synthesized and real contours.

It is well known that listeners are more sensitive to some parts of F0 contours then others, for instance listeners can perceive differences in peak height more readily that valleys. Unfortunately there is no known comparison technique that can mimic this behaviour and so we are forced to use a cruder approach whereby all parts of the contour are treated equally. To measure F0 contour similarity we use root-mean-squared error and correlation, which are somewhat standard in the literature (e.g. (Dusterhoff and Black, 1997), (Fujisaki and Ohno, 1997), (Ross and Ostendorf, 1994)).

Accuracy experiments were conducted on the 1061 utterances in the DCIEM test set. To obtain the Tilt and RFC representations for testing, we used the automatic analysis procedure described in section 5.1. This analysis was performed on the events derived from the hand transcriptions and on the events found by the automatic event detection process. For each utterance, the original raw and smoothed F0 contours were compared with the contours generated from the RFC and Tilt representations. Table 6.1 shows the rms error and correlation (ρ) for the comparisons.

Representation	**raw F0 rmse**	**raw F0 ρ**	**smooth F0 rmse**	**ρ smooth F0**
complete rfc	14.60	0.651	6.94	0.837
complete tilt	14.58	0.647	7.14	0.829
event only rfc	12.86	0.630	6.82	0.798
event only tilt	13.13	0.620	7.15	0.786
automatic rfc	15.11	0.651	7.16	0.841
automatic tilt	15.25	0.644	7.51	0.833

Table 8: Accuracy figures for RFC and Tilt synthesis

Two clear patterns emerge from the table. Firstly, looking at the first two rows of the table, we see that the artificial contours match the smoothed contours much better than the raw contours. The smoothing technique eliminates F0 tracking errors and segmental perturbations, so it can be argued that the smoothed contours are a more meaningful representation to measure against than the raw contours. In a simple attempt to focus on the relevant parts of the contour, the F0 regions within the event boundaries were subjected to the same analysis. Rows 3 and 4 show the results. There are slight improvements in rms error and slight reductions in correlation, but the overall pattern is the same as the errors for the complete contour. Rows 5 and 6 show the results for the contours synthesized from automatically detected events. The correlations are about the same and the rms errors are slightly worse than for the hand detected events.

The second pattern we find is that although RFC contours are closer to originals than Tilt contours are to originals, the difference is very small and often insignificant. In other words, the advantage in being able to convert RFC representations into Tilt representations isn't at the expense of much synthesis accuracy. To demonstrate this point further, a comparison was conducted between the synthetic RFC contours and the synthetic Tilt contours. The rms error on hand detected events was 0.975 and the correlation was 0.992, on automatically detected events the rms error was 1.26 and the correlation 0.98. To all intents the contours are identical. From the comparison of synthesis accuracy of the Tilt and RFC we can conclude that there is very little information lost in the RFC to Tilt mapping process. While section 5.5 showed that the 3 parameters in the Tilt representation are quite independent, the synthesis result shows that the dimension reduction in RFC to Tilt does not throw away much information.

In section 1.2 we described one of the goals of an intonation model as being "wide coverage". In fact the test as to whether or not a model has wide coverage can actually be formulated in terms of a synthesis test. Taken independently from the other 2 goals, all "wide coverage" actually means is that the representation being used, in conjunction with its analysis and synthesis processes, is powerful enough to describe the data under examination. By "describe" we mean having the ability to code the original data without information loss. So to check the coverage capability of a representation, all we have to do is analyse it in terms of the representation, synthesis the data from this and compare to the original. Representations giving low errors have wide coverage. Of course it is relatively easy to produce an arbitrary coding scheme which can do this if one doesn't pay attention to the other goals. But so far we *have* shown that the tilt model has satisfied goals 1, 4 and 5, namely that is uses a constrained representation, and is capable of automatic synthesis and analysis. We now turn to a discussion of the model's fulfilment of the final goal, namely linguistic meaningfulness.

7 Linguistic Meaning

We now turn to the question "is the Tilt representation linguistically meaningful?". The most frequent criticism that has been made about the RFC and Tilt models is they are "only a coding of the F0 contour, and aren't linguistically meaningful". A strict definition of the term linguistically meaningful has been avoided until now, because it is a complex issue and requires discussion in the light of the results reported above. It is difficult to come up with firm (i.e. experimental) evidence for the model's linguistic meaningfulness as there are no simple measures of this, in contrast to the fairly straightforward methods used to demonstrate the Tilt model's success at the other stated goals. However, when we look at this issue more thoroughly, it is clear that it is very difficult to justify the linguistic relevance of *any* existing model of intonation. The following sections discuss various aspects of this issue.

7.1 Applications

We have stated that the Tilt model has been designed to facilitate intonational processing for speech technology applications, and hence in the first instance we should address the concept of "linguistically meaningful" in this sense. For applications, the main requirement is that the Tilt representation is "usable". In a speech analysis environment (e.g. a speech recognition system), this means that Tilt representations should be interpretable by other system components which need to use intonational information. In a synthesis environment (e.g. a text-to-speech system) the requirement is that high level modules in the system can generate Tilt parameters from other linguistic representations.

The model has been used as the last component in several TTS intonation models. In Taylor (Taylor and Black, 1994) normalised (speaker independent) tilt parameters were generated from rule based feature descriptions. Speaker specific parameters were then used to produce normal tilt representations from which F0 contours were generated. Black (Black, 1997) describes a method for learning Tilt parameters automatically from data and then generating them at synthesis run time. In an extension of this work, Dusterhoff and Black (Dusterhoff and Black, 1997) describe a method for using CART to generate F0 contours from high level information in a text-to-speech system. They do this by training the decision trees to produce Tilt parameters. This study is particularly interesting in that they perform a direct comparison between the Tilt representation and a ToBI labelling of the same data. The Tilt representation gave slightly better performance, showing that at least in this setup it is useful representation.

The Tilt model has also been successfully applied to speech recognition. Wright and Taylor (Wright and Taylor, 1997) describe a system for automatically recognising the dialogue act of an utterance from an analysis of its intonation. Each utterance is automatically analysed using the tilt model and a HMM classifier is used to assign it to one of 12 dialogue act types (such as acknowledgement, yes-no question). This classifier has been used as a component in a speech recognition system and has been shown to help reduce word error rate (**?**).

7.2 The Bias Against Continuous Representations

The Tilt representation uses continuous variables to describe pitch accents, unlike more traditional representations which use discrete categories. Here we take some time to argue that continuous variables can legitimately form part of a linguistic intonational description.

It is traditional in linguistics to deal with categorical (i.e. discrete) representations alone, to such an extent that continuous representations are often deemed unlinguistic in some sense. This has led to properties of intonation that are clearly continuous, such as pitch range and prominence, being somewhat ignored, and study concentrating on categorical issues only, such as pitch accent type. Despite virtually overwhelming evidence that prominence and pitch range follow regular patterns and have an important linguistic function (Gussenhoven and Rietveld, 1988), (Terken, 1991), (Ladd, 1996), (Ladd, 1994), these parts of intonation are often called “paralinguistic” and omitted from intonational representations simply because of their continuous nature. The bias towards purely discrete representations in linguistics is a hangover from traditional linguistics and has often been justified because such representations are seen as being properly “cognitive” (**?**). Massaro (**?**) questions the whole basis of categorical perception in linguistics, explaining that the dominance of this idea arises from equating discrimination with identification, an equivalence which does not really hold. However, in recent years there has been a growing acceptance of continuous representations in linguistics, partly through the acceptance of connectionist models as legitimate cognitive science. As such the corollary that only discrete phenomena can be considered cognitive does not hold.

The proponents of what Ladd (Ladd, 1996) terms the Autosegmental-Metrical (AM) school of intonation (Pierrehumbert, 1980), (Liberman, 1975), (Bruce, 1977) have argued strongly that intonation has a phonological level of representation, in the same way as segmental phonology/phonetics does, and that the sounds patterns of intonation are best described with such representations. Much of the evidence for the phonological level has stemmed from showing that this solved many problems with previous approaches because with a phonology sound patterns can be described in abstract ways, without having to deal with pitch values directly. This allows meaningful comparison of intonation across speakers who have different pitch ranges, for instance. While we believe that this is one of the major contributions to modern intonation research, we believe the success of this approach is due to the adoption of an *abstract* level of representation in the broad sense, rather than a necessarily *phonological* one in the traditional discrete sense. Just because abstract representations have proved useful, and because they share many similarities with the structures of segmental phonology, it does not follow that abstract intonational representations are necessarily *phonological in the same way*.

A crucial property of segmental phonology is that the connection between the semantic and phonological properties of a lexical item is arbitrary. This is proved by showing that two words (for example "pill" and "bill") which sound similar do not necessarily mean similar things. Crucially, we can prove the segmental phonological space is discrete by showing that is impossible to *perceive* a sound which is half way between a /b/ or a /p/. It has been shown in synthesis experiments, that if listeners are played a pattern of words such as "pill", each time with more voicing in the initial stop, at some stage they will start to perceive the word "bill". At no point however, although the sound pattern is half way between a normal /p/ and /b/, will the listeners conjure up a halfway semantic image of an entity which is a bit "bill"-like and a bit "pill"-like. Thus although there might be an acoustic continuum, there is a sharp perceptual boundary which prohibits interim semantic representations.

There is no evidence that intonation behaves in such a way. It is clear that different acoustic intonation patterns can give rise to different semantic interpretations, the crucial point is that the connection between intonational sound and meaning is not arbitrary in the same way, and that if intonational sound S_A gives rise to meaning M_A and sound S_B gives rise to meaning M_B, then a sound halfway between S_A and S_B can certainly give rise to a meaning somewhere between M_A and M_B. In other words there has been no evidence to show that there are strict boundaries between intonational units which signal abrupt changes in meaning.

Ladd (Ladd, 1996) in fact uses just this argument as support for there being proper, discrete phonological categories in intonation. In linguistics (including the phonological part of intonation) he claims that "close similarity of phonetic form is generally of no relevance for meaning" and states that in contrast "semantic continua are matched by phonetic ones" in paralinguistics. Crucially, he does not state the *actual* defining properties of a categorical system, namely that there should be strict and identifiable boundaries between the categories.

Further evidence for problems with categorical intonational classification come from consideration of how one would actually go about producing a phonological inventory for a new accent or language. In segmental linguistics, the classical way to determine the phonological units of a language is via the use of minimal pairs. This is how one can find out whether for instance [r] and [l] are distinct phonological units (as in English) or whether they are allophonic (as in Japanese). By knowing that the words "crown" and "clown" have clearly different meanings, we know that [r] and [l] are phonologically distinct units. No-one has yet produced an equivalent test for intonational units.

Traditionally the argument about categorisation in intonation has revolved about a false dichotomy, namely that the relationship between sound and meaning, can either be as in segmental phonology, where the relationship is completely arbitrary, or as in paralanguage, where the relationship is a simple linear one. In fact this is an inappropriate application of the law of the excluded middle and one does not have to choose either of these positions: a third position is that intonation is continuous with regard to both sound and meaning, but that the relationship between the two is highly complex and non-linear. Adopting such a position can explain why simple attempts to prove direct correspondence between sound and meaning in intonation have failed, but also why it is so hard to produce evidence for categorical boundaries. In this view pitch accents occupy positions in a multi-dimensional sound space, and in effect what H* and L* etc represent are points of particular importance in this space. One can think of this as somewhat analogous to how people describe the temperature of an object when they are touching it. Physically, temperature is a continuum with no distinct categories, but it is helpful to have terms such as hot and cold which describe certain temperature situations. This is not troublesome so long we accept this as just a convention, and we don't insist that underlying temperature is categorical. It is pointless to go further try and define strict boundaries on what is underlyingly a continuous phenomena. While there will be a lot of agreement as to what hot and cold represent under these conditions, there will always be temperatures between the two which are impossible to categorise either way. Going back to intonation, it is clear that a typical H* accent is different from a typical L* accent, the point is that there are accents in between which could be described as either. We again re-iterate that point that proof of the existence of categories depends on the proof of the existence of category boundaries.

7.3 Phonetics and Phonology in the Tilt Model

The Tilt representation as described here can be termed *phonetic* because its purpose is to describe observable linguistic sound phenomena. Although the focus of our paper is on this representation and its relationship with the acoustics, it is useful to informally discuss the relationship between the phonetic Tilt representation and higher level, phonological representations. The parameters used to describe events in the Tilt representation are quite literal with respect to measurable acoustic quantities, and hence we have duration and position measure in seconds, and amplitude in Hertz. We advocate that a phonological representation in the Tilt model should have the same parameters as the phonetic representation (with the possible exception of duration, see below), but that their *scales* should be modified so as to represent higher level phenomena more appropriately.

For instance, the amplitude parameter should reflect genuine perceptual prominence, rather than simple acoustic magnitude. Different speakers have different pitch ranges and these differences should be accounted for in a phonological representation. It has been widely shown (Cohen et al., 1982), (Pierrehumbert, 1980), (Liberman and Pierrehumbert, 1984), (Ladd, 1984) that pitch range narrows towards the end of an utterance. This means that a pitch accent at the start of an utterance needs a bigger F0 amplitude than an accent at the end to produce an effect of equal prominence in a listener. Hence it would be desirable for phonological Tilt amplitude to be normalised with respect to pitch range and that amplitude should be a measure of perceived prominence.

As stated in section 5.4, the duration and position parameters are dependent on the local segmental content of the utterance which is undesirable from a phonological point of view. In fact it is possible that the duration parameter is wholly dependent on segmental content and carries no phonological information at all. If this were the case, then the number of parameters in the phonological Tilt representation could be reduced. The position parameter, which certainly does contain high level information, should be normalised with respect to the segmental content.

The tilt parameter itself is dimensionless and so is independent of amplitude and time scales. Hence it is possible that this is already as abstract as it needs to be and would not require modification.

It light of our previous discussion we think it is entirely appropriate for the phonological representation to have continuous parameters. The key point about the scales of phonological representation is that events which are perceived as being the same should have the same values in the Tilt representation.

7.4 Comparing the Tilt and AM/ToBI models

We now discuss the similarities between the tilt model and the AM/ToBI models. In many respects the models are very similar. Both adopt the same approach to intonational primitives, namely that the intonational representation of an utterance should be a linear sequence of event based intonational entities, associated with syllables/segments in an autosegmental structure. Following from this, both agree that downdrift in intonation can be accounted for by a combination of pitch accent downstepping and gradual falls in pitch range, and hence that downdrift is not modelling by global phrase patterns. While the phonetic aspects of the AM are often ignored, both models agree that intonation can be described using

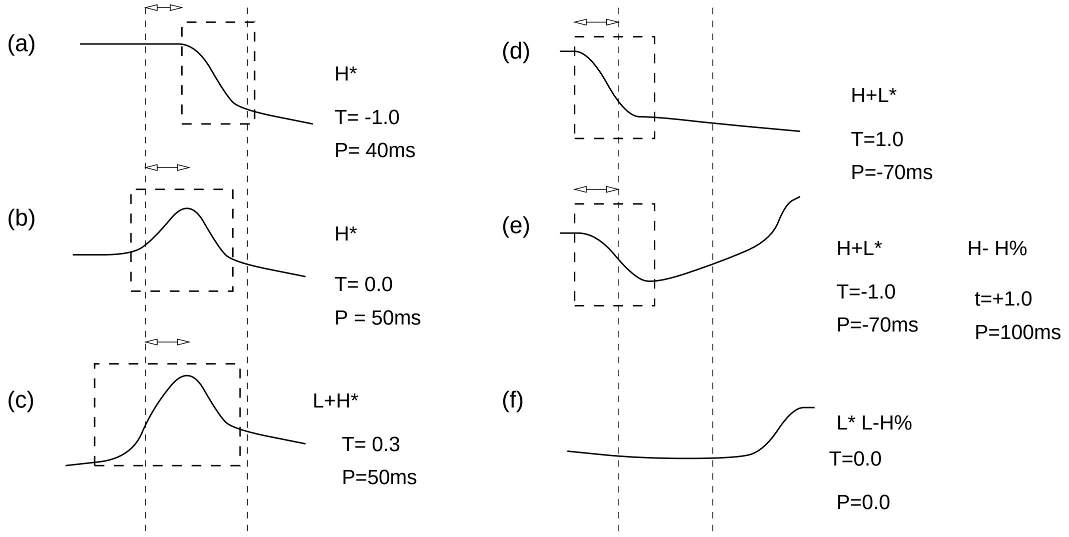

Figure 4: Examples of Tilt and AM representations for various common types of utterance

a high level abstract sound representation, i.e. a phonology.

Some differences arise from the phonetics/phonology mismatch of the models. Phrase tones (H- and L-) are used in the AM model as a mechanism to differentiate four types of post-nuclear intonation. Phrase tones are purely phonological units, having no direct F0 realisation and hence there is no equivalent for these in the Tilt model. In a similar way, there is no equivalent for the low boundary tone (L%). Falling post nuclear intonation is usually modelled by a single connection, and only rising boundaries have an event. The observed variation in post nuclear intonation is modelled by using different parameters for the connections and rising events. Level accents may legitimately be part of a phonological representation but are absent from the phonetic one because they have no observable F0 behaviour. If deemed desirable, such accents could be accommodated in the tilt model as entities with zero amplitude, duration and tilt, and hence maybe the issue as the whether or not level accents should be represented is not of great consequence.

Accepting that Tilt duration is probably a purely phonetic phenomena, we now discuss the relationship of the more phonological tilt parameters (amplitude, tilt and position) with respect to the AM model. To simplify the discussion, we assume that these three parameters have been normalised and are represented on phonological scales in the way described in section 7.3. Some varieties of the AM model, for example the original Pierrehumbert system (Pierrehumbert, 1980), (Ladd, 1987), actually have an amplitude parameter very similar to the type we propose for the phonological tilt model. Because of the paralinguistic argument, this is often treated separately from the system used to describe accent type. Having accounted for duration and amplitude, we now turn to showing how the the position and tilt parameters relate to the tonal accent classification system of the AM model. Figure 6 shows typical tilt and position parameters for some accent and boundary tone combinations. Figure 7 is an impressionistic plot, with one axis representing tilt and the other position. This plot maps the space of possible pitch accents. The relevant AM accents and boundary tones have been marked to show how each relates to the tilt parameters.

As should be clear by now, the main area in where the models disagree is that the Tilt representation uses continuous parameters to model events whereas the AM model uses discrete classes. Although we think that it is an important point in its own right to demonstrate the tenuous nature of categorical descriptions in intonation, we further argue that the adoption of such description mechanisms is actually

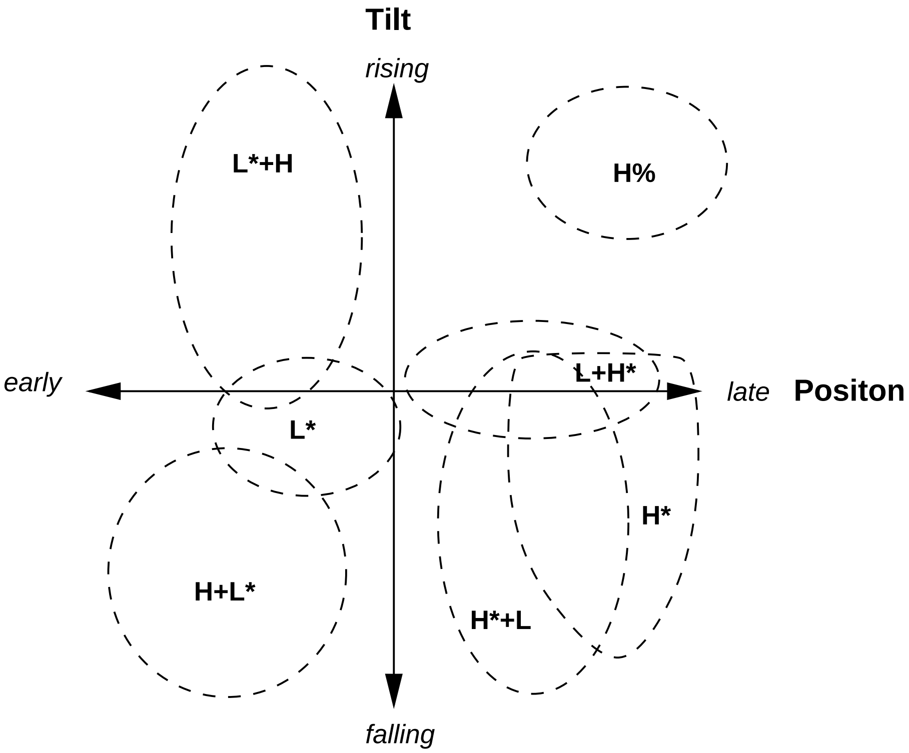

Figure 5: Two dimensional schematic representation of intonational space

harmful in that it doesn't provide a satisfactory means to describe intonational events. The first problem is that because of the difficulty in defining class boundaries, it is very difficult for machines and humans to label naturally occurring speech using the AM system. While it is simple enough for experts to produce canonical examples of different pitch accents and get agreement when labelling these, the situation is very different in natural speech. Leaving the distribution problem aside (see below), it has been our experience that even linguists trained in intonation find it difficult to decide which class a particular accent belongs to. As human labellers must provide the data on which automatic systems are to be trained and tested, badly labelled data cannot be expected to facilitate accurate automatic systems. The second problem is that in naturally occurring speech, the distribution of AM pitch accent types is extremely uneven. In the RN corpus about 79% of all accents in the corpus fall into the H* category, 15% into the L+H* category. From an information theoretic point of view, any classification system which lumps the vast majority of tokens into a single type isn't very useful because it doesn't actually provide tell us much about the tokens. Importantly, within the vast H* class, there actually is a substantial amount of variation which is linguistically meaningful and this information is lost.

Theses problems of labelling difficulty and unevenness of distribution are not present in the Tilt model as the continuous parameters express the differences in pitch accents naturally without having to resort to a forced classification.

7.5 Comparing the Tilt and Other models

The Fujisaki model (Fujisaki and Ohno, 1997) has many similarities to the Tilt model. It is a fully formal and quantitative phonetic model and hence has a defined synthesis algorithm. It too models accents as events, with no categorical accent types built in to the model and it uses an amplitude and duration parameter for event parameterisation.

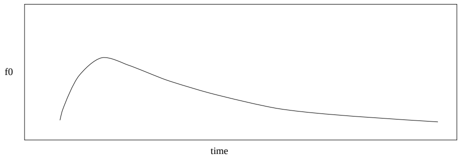

Figure 6: Fujisaki phrase component

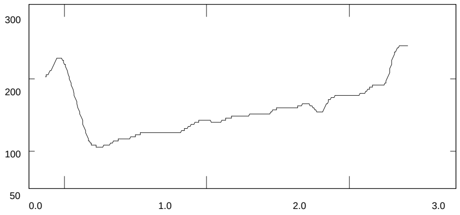

Figure 7: Contour with slowly rising intonation

However, the models differ in significant ways. Firstly, the tilt model allows the gradients of the rise and fall parts of events to vary. While some gradient variation in the rises and falls of the Fujisaki models accents is possible, this is minor and indirect and due to the gradient of the underlying phrase component. From our study of rise and fall pitch accent gradients in our database, it is clear that substantial variation is possible, and that any model which advocates fixed gradients will have substantially worse synthesis accuracy that the Tilt model. Secondly, the Fujisaki model makes use of a phrase component, which dictates the global shape of the F0 contour over the length of a phrase. The shape of this is shown in figure 6. Again, we think this is too constrained and can not account for the wide range of observed long term contour shapes in English. For example in the contour shown in figure 7, the intonation is slowly and steadily rising after the first event. Fujisaki (personal communication) has suggested that this can be modelled by stacking several phrase components on top of each other in short intervals. While such a solution could recreate the shape shown in the contour, it is has a severe price, in that all reference between the position of the phrase component and its linguistic meaningfulness has been lost.

In summary, it seems the Fujisaki model does not have enough degrees of freedom to synthesize English intonation and still keep some sense of linguistic relevance.

8 Conclusion

We conclude by an examination of how well the Tilt model has fulfilled the five goals described in section 1.2. The issues of *constraint* and *wide coverage* go hand in hand in they are essentially concern the *degrees of freedom* of a model. Put simply, a model with too many degrees of freedom will be able to model the data well but will have a large amount of redundancy in its representation. A model with too few degrees of freedom will have a compact representation but will not be able to synthesize the data accurately nor differentiate certain phenomena. The results in table 5.5 show that the Tilt representation is compact and has low redundancy, while still producing wide coverage, as shown by the synthesis accuracy results in table 6.1. We therefore conclude that an appropriate balance has been found between constraint and wide coverage. We have stated that the duration parameter may be purely phonetic in nature, and that only amplitude, position and tilt should be considered as phonological parameters.

The proof of any representation's linguistic meaningfulness if difficult, but we hope that the discussion of this issue in section 7 has at least shown that continuous parameters can form the basis of a high level intonational representation and that furthermore, the tilt and position parameters map the intonational space elegantly. In the end we take a pragmatic approach to the issue of deciding whether one representation system is better than another. If the bias against continuous representations is removed, we think that the Tilt model compares well against the AM model and in that it provides a simple representation which solves many of the problems that categorical representations impose.

A long raging argument in the field has been whether intonation should be regarded as a tonal or contour (shape) phenomena (see (Ladd, 1996)). While this may be a interesting theoretical question, for the computational goals presented here, it is somewhat beside the point. For practical purposes, a more important goal than discovering the nature of intonational primitives is to assess a model in terms of the criteria explained in the introduction. A model with good coverage, accuracy and linguistic meaningfulness is better than a model without these, regardless of which one is tonal or contour.

It is in the area of automatic analysis that perhaps most future improvement could be made. Spontaneous speaker independent conversational speech is one of the most challenging types of speech for any system, and hence it is a solid achievement that the event detector can achieve the level of success that it does on this task. However, with performance figures on the DCIEM data of 72.7% correct, 47.7% accuracy for all accents and 81.9% and 60.7% for non-minor accents there is obviously room for improvement. It should be pointed out that it is not necessary to use this particular event detection algorithm with the Tilt model: any system that can locate events and give their boundaries can be used. As explained in section 4.1 our system is an event *detector* but there is nothing to stop an event *classifier* (e.g. (Ross and Ostendorf, 1995)) being used in conjunction with the model. The choice of a classifier or detector essentially depends on whether a phonetic segmentation is available in the application.

Table 6.1 shows that the Tilt model can synthesize F0 contours with a high degree of accuracy. Given the just noticeable difference for F0 in natural speech is about 4 Hz (Hess, 1983), the figures for the comparison between the synthesized and smoothed contours show that the synthetic contours are nearly identical to the smoothed ones. Raw contours differ from smoothed ones mainly due to the presence segmental glitches and perturbations. The Tilt synthesis procedure has not attempted to model these is any way, and hence the errors for the synthetic versus raw comparison is worse.

In summary, we have shown that the Tilt model has been fairly successful at fulfilling our original goals. We have tested the model rigorously on read speech and spontaneous dialogue speech from a number of different speakers in different situations. We think the success of the model has been to find a suitable balance between the conflicting goals, so that the model facilitates automatic analysis and synthesis and provides a useful and elegant linguistic representation of intonation.

Appendix A: Mapping from ToBI to Tilt labels

Despite the statement in section 7.4 that the Tilt and AM models are both based on intonational events, there are important technical differences in the way that AM (i.e. ToBI) and Tilt label files are actually constructed. In section 4.5 we described experiments that a Tilt transcription of the Boston University Radio News (RN) corpus was derived by from the original ToBI transcriptions. We briefly explain how this was done.

The RN corpus was labelled using the guidelines in Beckman and Ayers (Beckman and Ayers, 1993) and the transcriptions are formatted in xwaves xlabel files. Xwaves files contain a list of elements, one per line, where each line contains a label name and a time, normally representing the end of the label. In the ToBI files, the time in each line represents the notional *center* of the pitch accent, e.g. the time of an H* is marked at its peak. As the Tilt event detector is trained on events whose start and end times are marked, the ToBI transcriptions need to be realigned so as accent start and end times rather than accent middles are marked. There is no single "right" way, but we estimate the start and stop times by assuming that the time given is the literal middle of the event, and then marking the start and end times as being $d/2$ before and after this where d is the average duration of an event in another database (d=210ms in the DCIEM corpus). Connections are inserted between non-contiguous events and phrase tones and low boundary tones are deleted.

Appendix B: Software and Corpora Availability

The HTK toolkit was used in the event detection experiments. It is available under licence from Cambridge Entropic Labs, U.K (Young et al., 1996).

All the other software, including the super resolution pitch detection algorithm, Tilt analysis, Tilt synthesis and transcription comparison code is included in the Edinburgh Speech Tools, a publicly available speech software toolkit. This is part of the Festival speech synthesis system available from http://www.cstr.ed.ac.uk/projects/festival.

The CART based F0 generation algorithm (Dusterhoff and Black, 1997) which uses the Tilt model is implemented in the Festival speech synthesis system, available from the above address.

The original DCIEM, Boston University Radio News and Switchboard databases can be licenced from the Linguistic Data Consortium, http://www.ldc.upenn.edu. These databases contain the original waveforms and transcriptions.

Derived F0 and energy contours, and the Tilt model labellings for all the experiments reported here are available from http://www.cstr.ed.ac.uk/projects/intonation.

Acknowledgements

References

Bagshaw, P. C., Hiller, S. M., and Jack, M. A. (1993). Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *Proc. Eurospeech '93, Berlin*.

Bard, E. G., Sotillo, C., Anderson, A. H., and Taylor, M. M. (1995). The DCIEM map task corpus: Spontaneous dialogues under sleep deprivation and drug treatment. In *Proc. of the ESCA-NATO Tutorial and Workshop on Speech under Stress, Lisbon*.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. *Inequalities*, 3:1–8.

Beckman, M. E. and Ayers, G. M. (1993). Guidlines for ToBI labelling. Technical Report Version 1.5, Ohio State University.

Black, A. W. (1997). Predicting the intonation of discource segments from examples in dialogue speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody*. Springer.

Bruce, G. (1977). *Swedish Word Accents in Sentence Perspective*. PhD thesis, University of Lund.

Cohen, A., Collier, R., and t'Hart, J. (1982). Declination: construct or intrinsic feature of speech pitch. *Phonetica*, 39:254–73.

Dusterhoff, K. and Black, A. (1997). Generating intonation contours for speech synthesis using the tilt intonation theory. In *ESCA workshop on Intonation: Theory Models and Applications*.

Fujisaki, H. and Ohno, S. (1997). Comparison and assessment of models in the study of fundamental frequency contours of speech. In *ESCA workshop on Intonation: Theory Models and Applications*.

Gauvain, J., Lamel, L., Adda, G., and Matrouf, D. (1996). Developments in continuous speech dictation using the 1995 arpa nab news task. In *International Conference on Speech and Signal Processing*. IEEE.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP92*, pages 517–520.

Granstrom, B. (1997). Applications of intonation - an overview. In *ESCA workshop on Intonation: Theory Models and Applications*.

Gussenhoven, C. and Rietveld, T. (1988). Fundamental frequency declination in Dutch: Testing three hypotheses. *Journal of Phonetics*.

Hess, W. (1983). *Pitch Determination of Speech Signals*. Springer-Verlag.

Hirst, D. (1992). Prediction of prosody: An overview. In Bailey, G. and Benoit, C., editors, *Talking Machines*. North Holland.

Ladd, D. R. (1984). Declination: A review and some hypotheses. *Phonlogy Yearbook 1*.

Ladd, D. R. (1987). A model of intonational phonology for use with speech synthesis by rule. In *European Conference on Speech Technology*. ESCA.

Ladd, D. R. (1994). Constraints on the gradient variability of pitch range. *Papers in Laboratory Phonology 3*, pages 43–63.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge Studies in Linguistics. Cambridge University Press.

Lea, W. A. (1980). *Prosodic Aids to Speech Recognition*. Prentice Hall.

Liberman, M. (1975). *The Intonational System of English*. PhD thesis, MIT. Published by Indiana University Linguistics Club.

Liberman, M. and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In Aronoff, M. and Oehrle, R. T., editors, *Language Sound Structure*. MIT Press.

Medan, Y., Yair, E., and Chazan, D. (1991). Super resolution pitch determination of speech signals. *IEEE Trans. Signal Processing*, 39:40–48.

Ostendorf, M., Price, P. J., and Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus. Technical Report ECS-95-001, Boston University.

Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT. Published by Indiana University Linguistics Club.

Ross, K. and Ostendorf, M. (1994). A dynamical system model for generating F0 for synthesis. In *Second ESCA/IEEE Workshop on Speech Synthesis, New York*.

Ross, K. and Ostendorf, M. (1995). A dynamical system model for recognising intonation patterns. In *EUROSPEECH 95*, pages 993–996.

Secrest, B. G. and Doddington, G. R. (1993). An integrated pitch tracking algroithm for speech systems. In *ICASSP 83*, pages 1352–1355.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP92*, volume 2, pages 867–870.

Taylor, P. A. (1995). The rise/fall/connection model of intonation. *Speech Communication*, 15:169–186.

Taylor, P. A. and Black, A. W. (1994). Synthesizing conversational intonation from a linguistically rich input. In *Second ESCA/IEEE Workshop on Speech Synthesis, New York*.

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, 89:1768–76.

t'Hart, J. and Collier, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics*, 3:235–255.

Waibel, A. (1986). *Prosody in Speech Recognition*. PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Waibel, A., Finke, M., Gates, D., M. Gavalda, T. K., Lavie, A., Levin, L., Maier, M., Mayfield, L., McNari, A., rogine, I., Shima, K., sloboda, T., Woszczyna, M., Zeppenfeld, T., and Zhan, P. (1996). Janus-II - advances in spontaneous speech recognition. In *International Conference on Speech and Signal Processing*. IEEE.

Woodland, P., Leggetter, C. J., Odell, J. J., Valtcher, V., and Young, S. J. (1995). The 1994 HTK large vocabulary speech recognition system. In *International Conference on Speech and Signal Processing*. IEEE.

Wright, H. and Taylor, P. A. (1997). Modelling intonational structure using hidden markov models. In *ESCA workshop on Intonation: Theory Models and Applications*.

Young, S., Jansen, J., Odell, J., Ollason, D., and Woodland, P. (1996). *HTK manual*. Entropic.